D302–D305 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 3 November 2011*
doi:10.1093/nar/gkr931

SMART 7: recent updates to the protein domain annotation resource

Ivica Letunic, Tobias Doerks and Peer Bork*

EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany

Received October 5, 2011; Accepted October 10, 2011

ABSTRACT

SMART (Simple Modular Architecture Research Tool) is an online resource (http://smart.embl.de/) for the identification and annotation of protein domains and the analysis of protein domain architectures. SMART version 7 contains manually curated models for 1009 protein domains, 200 more than in the previous version. The current release introduces several novel features and a streamlined user interface resulting in a faster and more comfortable workflow. The underlying protein databases were greatly expanded, resulting in a 2-fold increase in number of annotated domains and features. The database of completely sequenced genomes now includes 1133 species, compared to 630 in the previous release. Domain architecture analysis results can now be exported and visualized through the iTOL phylogenetic tree viewer. 'metaSMART' was introduced as a novel subresource dedicated to the exploration and analysis of domain architectures in various metagenomics data sets. An advanced full text search engine was implemented, covering the complete annotations for SMART and Pfam domains, as well as the complete set of protein descriptions, allowing users to quickly find relevant information.

INTRODUCTION

The SMART database (http://smart.embl.de) is now in its 13th year (1), and provides high quality, manually curated Hidden–Markov models and alignments of protein domain families. Accessible though a web interface or via various programmatic methods, SMART remains a popular tool for domain annotation and exploration of protein domain architectures, with an average of 200 000 user submitted proteins analyzed monthly.

IMPROVED DOMAIN COVERAGE

Even though the rate of novel domain discovery is constantly declining (2), SMART gradually expands its domain coverage in each release. The current version 7 introduces more than 200 new domains, bringing the total to 1009 distinct modules that can be searched. Even though many of these domains were already annotated in other databases, like Pfam (3), SMART's domain annotation pipeline relies heavily on manual intervention, making the re-annotation process worthwhile.

UPDATED PROTEIN DATABASES

The number of annotated protein sequences is constantly growing, at the same time increasing the redundancy in the databases. Since protein redundancy significantly skews the number of domains reported in both domain architecture analyses and when comparing domain counts in complete genomes, past versions of SMART (4) introduced several features to minimize these problems. The standard protein database used by SMART combines the complete Uniprot protein database (5) with predicted proteins from all stable Ensembl (6) genomes. Since these are inherently highly redundant, SMART implements a per-species clustering method (7) to minimize the redundancy in the final database. Yet, the updated version currently contains more than 11 million proteins from around 150 thousand species, subspecies and varietas. Additionally, SMART offers a 'genomic' analysis mode that contains only proteins from completely sequenced genomes. Synchonized with STRING version 9 (8), this database has been significantly expanded, and contains 1133 complete genomes (121 Eukaryota, 943 Bacteria and 69 Archaea).

NOVEL ARCHITECTURE ANALYSIS DATA EXPORT AND VISUALIZATION FEATURES

Domain architecture analysis functions in SMART allow users to simply access proteins containing combinations of particular domains. These can be also generated using

*To whom correspondence should be addressed. Tel: +49 6221 387 8526; Fax: +49 6221 387 517; Email: bork@embl.de

combinations of GO terms (9) associated to protein domains, and restricted to various taxonomic classes. Previous versions of SMART allowed users to download these selected proteins as FASTA formatted files or to display them through schematic representations (SMART 'bubblograms'). SMART 7 offers a new data export functions for domain architecture analysis, which is tightly coupled with iTOL [interactive Tree Of Life (10,11)], our phylogenetic tree visualization tool.

Data are exported into two separate files, which can be directly used by iTOL: a Newick formatted phylogenetic tree and a protein domain data set file used to visualize proteins on the tree. The procedure is as follows:

(1) an initial list of proteins is obtained through an architecture analysis query;
(2) proteins are grouped according to their species of origin;
(3) these species are used to 'prune' the complete NCBI taxonomy database (12) by walking the taxonomy tree up to the root and exporting the resulting structure into a Newick formatted phylogenetic tree; and
(4) each protein's domain organization is converted into a plain text format understood by iTOL.

Resulting plain text files can be downloaded, or directly visualized in iTOL by a simple button click (Figure 1).

EXPANDED PROTEIN INTERACTION DATA

Similar to previous SMART updates, we synchronized our underlying protein interaction data with the latest version of the STRING database (8). Since the number of species in our protein database based on completely sequenced genomes increased almost 2-fold in this release, the information on putative protein interaction partners has also been significantly expanded, and is now available for more than 3.5 million proteins. Interaction network data display has been updated, and uses a streamlined graphical representation, which brings several extra layers of information while being easier to interpret.

metaSMART: BASIC INTEGRATION OF ENVIRONMENTAL SEQUENCING DATA

Metagenomics projects (that is environmental shotgun sequencing) are constantly increasing the amount of novel, uncharacterized DNA and (fragments of) protein sequences. Functional characterization and annotation of such data remains a daunting task, and various pipelines, such as SmashCommunity (13), are being developed to help scientists in this process.

As an initial step toward meaningful integration of these data into SMART, we created 'metaSMART'. Its primary goal is the exploration and analysis of protein domain architectures in various publicly available metagenomics data sets.

Users can compare different domain frequencies, co-occurrences and complex architectures in different environments to illustrate the role of domain variability depending on the habitat. Furthermore, metaSMART allows the exploration of completely novel domain architectures, unique in databases so far; analyses of various non-described domain compositions could broaden the knowledge about new protein functions related to their domain interdependency (Figure 2). Four metagenomics data sets are the starting point of metaSMART: Sargasso sea (14), acid mine drainage biofilm (15), Minnesota farm soil (16) and 'Whale fall' carcasses (16). We are currently integrating several additional metagenomes [for example, the human gut (17)], which will significantly expand the amount of available information in metaSMART and provide novel biological insights in the context of metagenomics.

DATABASE AND WEB SERVER OPTIMIZATIONS

The backend of SMART is a PostgreSQL-based relational database management system, which stores the annotation of all SMART domains and the pre-calculated protein analyses for the entire Uniprot (18), Ensembl (19) and STRING (8) sequence databases. These include SMART and Pfam domains, as well as several protein intrinsic features, like signal peptides, transmembrane and coiled-coil regions. With close to 50 million annotated features in the current database, we have to constantly find new ways of keeping the response times of the server acceptable. Therefore, the database was restructured and several parts of the database access code have been optimized. Additionally, the hardware cluster that powers the sequence annotation searches and database queries has been refreshed and expanded with additional CPUs.

USER INTERFACE IMPROVEMENTS

Version 7 brings various updates to SMART's web interface. Many parts of the interface have been simplified and compacted, resulting in easier navigation and simpler identification of relevant content. To make SMART more accessible to new users, we added help popup windows to various parts of the interface, making different functions easier to understand.

A new full text search engine has been implemented, based on KinoSearch libraries (http://incubator.apache.org/lucy). It indexes the complete annotation pages for all SMART and Pfam domains, as well as Uniprot, Ensembl and STRING protein descriptions, allowing users to quickly identify domains or proteins of interest.

Programmatic access to SMART has been extended with easy to parse text-only output mode, allowing simple batch access to the SMART search engine. Ready to use example scripts that use the batch access interface are also provided.

FUNDING

EMBL (internal budget) and the European Union under the program 'FP7 capacities: Scientific Data Repositories' (grant 213037) (IMproving Protein Annotation and Co-ordination using Technology – IMPACT). Funding for open access charge: EMBL (internal budget).

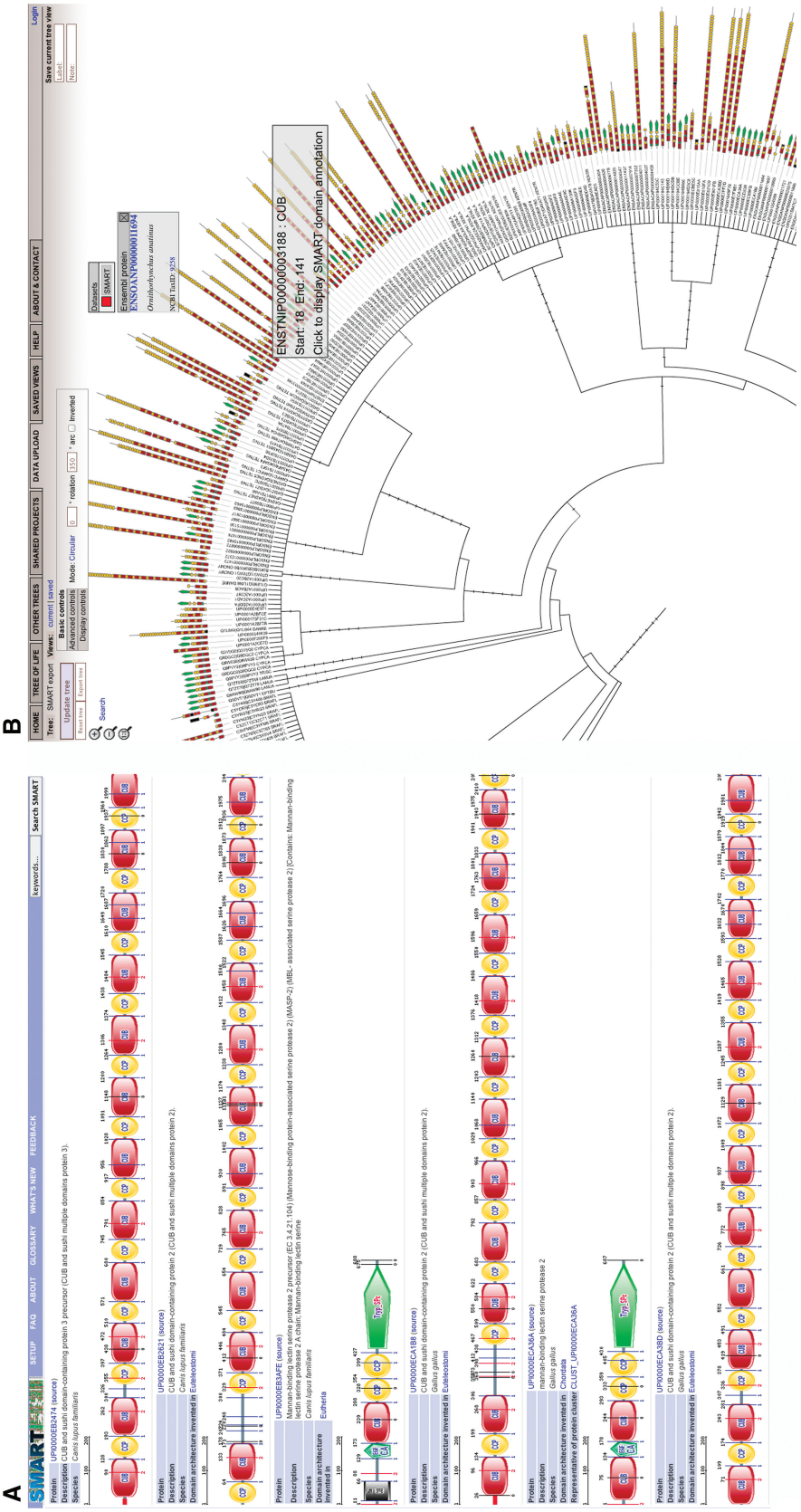

Figure 1. Displaying SMART protein domain architectures in iTOL. New data export features allow users to simply display domain architecture query results on a NCBI taxonomy based phylogenetic tree. Phylogenetic trees are generated on-the-fly by pruning the NCBI taxonomy database (12) and visualized in interactive Tree Of Life (10). (**a**) SMART was queried for all proteins containing both CUB and CCP domains. (**b**) Query results visualized on a phylogenetic tree in iTOL.

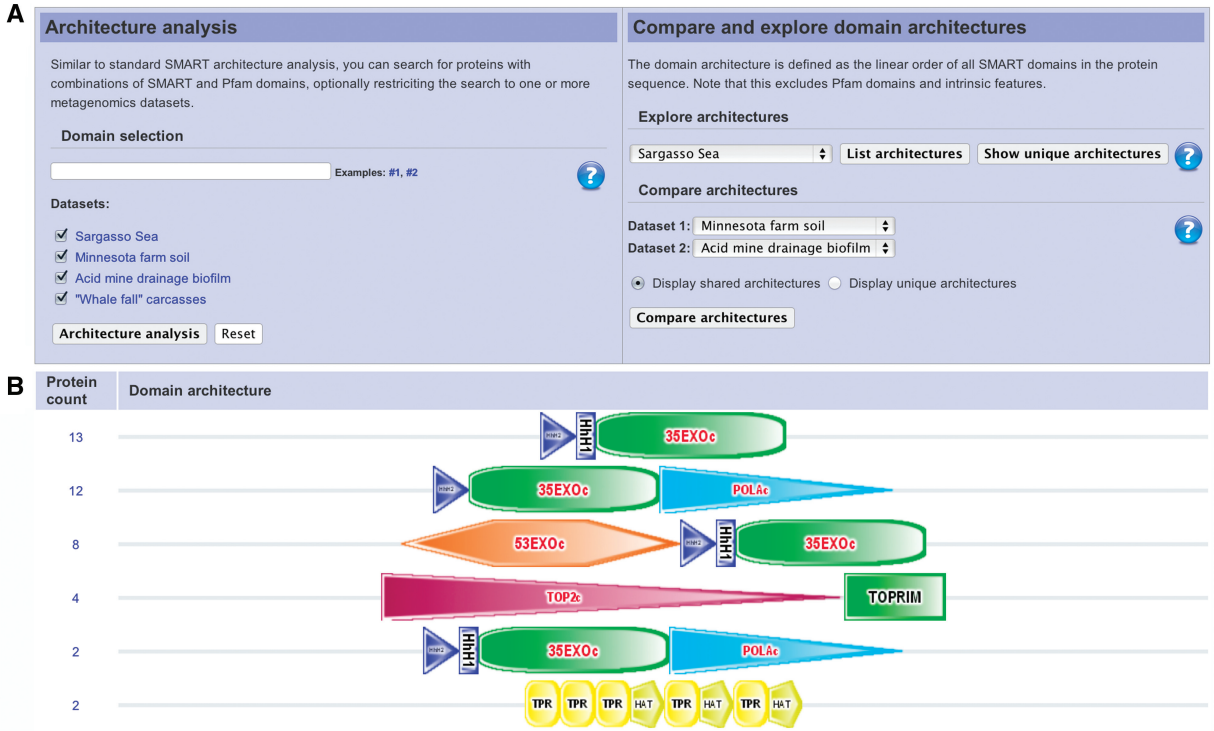

Figure 2. metaSMART, a novel sub resource dedicated to the exploration of domain architectures in metagenomics data sets. (**a**) metaSMART user interface provides simple access to all available functions. (**b**) A subset of protein domain architectures present in the Sargasso Sea data set (14). These are not present in other metagenomics data sets or the standard SMART database, and could be pointing to novel functional associations of various domains.

Conflict of interest statement. None declared.

REFERENCES

1. Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.
2. Heger,A. and Holm,L. (2003) Exhaustive enumeration of protein domain families. *J. Mol. Biol.*, **328**, 749–767.
3. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
4. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
5. Consortium,T.U. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
6. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
7. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
8. Szklarczyk,D., Franceschini,A., Kuhn,M., Simonovic,M., Roth,A., Minguez,P., Doerks,T., Stark,M., Muller,J., Bork,P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
9. Blake,J.A. and Harris,M.A. (2008) The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis. *Curr. Protoc. Bioinformatics*, Chapter 7, Unit 7 2.
10. Letunic,I. and Bork,P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.*, **39**, W475–W478.
11. Letunic,I. and Bork,P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127–128.
12. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Federhen,S. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.
13. Arumugam,M., Harrington,E.D., Foerstner,K.U., Raes,J. and Bork,P. (2010) SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics*, **26**, 2977–2978.
14. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
15. Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S. and Banfield,J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.
16. Tringe,S.G., von Mering,C., Kobayashi,A., Salamov,A.A., Chen,K., Chang,H.W., Podar,M., Short,J.M., Mathur,E.J., Detter,J.C. *et al.* (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554–557.
17. Arumugam,M., Raes,J., Pelletier,E., Le Paslier,D., Yamada,T., Mende,D.R., Fernandes,G.R., Tap,J., Bruls,T., Batto,J.M. *et al.* (2011) Enterotypes of the human gut microbiome. *Nature*, **473**, 174–180.
18. Consortium,T.U. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
19. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.